DNN Multimodal Fusion Techniques for Predicting Video Sentiment

DNN Multimodal Fusion Techniques for Predicting Video Sentiment

Jennifer Williams, Ramona Comanescu, Oana Radu, and Leimin Tian
Centre for Speech Technology Research (CSTR)
University of Edinburgh, UK
j.williams@ed.ac.uk

Abstract

We present our work on sentiment prediction using the benchmark MOSI dataset from the CMU-MultimodalDataSDK. Previous work on multimodal sentiment analysis have been focused on input-level feature fusion or decision-level fusion for multimodal fusion. Here, we propose an intermediate-level feature fusion, which merges weights from each modality (audio, video, and text) during training with subsequent additional training. Moreover, we tested principle component analysis (PCA) for feature selection. We found that applying PCA increases unimodal performance, and multimodal fusion outperforms unimodal models. Our experiments show that our proposed intermediate-level feature fusion outperforms other fusion techniques, and it achieves the best performance with an overall binary accuracy of 74.0% on video+text modalities. Our work also improves feature selection for unimodal sentiment analysis, while proposing a novel and effective multimodal fusion architecture for this task.

1 Introduction

Sentiment analysis is the study on the underlying attitude that one holds towards a certain entity. For a long time, text-based sentiment analysis has been the staple in this area and only recently are other modalities being considered for sentiment analysis such as vision and speech (Poria et al., 2015). For text channels, the features usually include information about word sequences and meaning (Mikolov et al., 2013). However, combining information from multiple modalities can bring additional information to ambiguous cases. For example, a smile extracted from facial features could help disambiguate cases such us "This movie is sick". Text alone would have trouble interpreting the meaning of the word "sick" in this context. This motivates the research of multimodal sentiment analysis. We seek to exploit the inter-dependencies between audio, text, and visual modalities in order to label video segments that exhibit positive or negative sentiment.

In current studies in this field, visual features often involve salient points of the face or body (Zadeh et al., 2016a), while low-level descriptors are collected from the speech signal such as pitch and volume (Zeng et al., 2009). The combination of features which have originated from text, speech and audio is what forms the basis of our multimodal classification work. Features from each modality are modeled, learned, and eventually *fused* together at various levels in a classification Deep Neural Network (DNN) system. When the modalities are fused together, this is called *multimodal fusion*. DNN multimodal fusion for binary sentiment classification is an active area of research that continues to gain momentum and spark interest due to the challenging nature of the problem (e.g., Poria et al. (2018)). We explore the interplay between three modalities: text, video, and audio. We focus on three fusion techniques inspired by previous work on multimodal fusion (Poria et al., 2018; Zadeh et al., 2016b).

We developed and compared three multimodal fusion architectures: (1) Input-level features fusion, (2) Intermediate features fusion, and (3) Decision-level fusion (late fusion). The first method refers to fusing information at the level of input features, similar to an unweighted concatenation of feature vectors, and it is the most widely used. The second method evokes the notion that each modality can be learned using a unimodal DNN. The weights learned through train-

ing each unimodal DNN are concatenated together and training continues before the decision level. The third method, also known as *ensemble fusion* or *late fusion*, fuses multiple modalities at the decision level. We present our multimodal DNN fusion approaches in detail in our methodology description in Section 3. where we further analyze the interactions between modalities. We experimented with combinations of modalities as well as system architectures that attempt to capture the interplay between modalities.

2 Related Work

Sentiment analysis has traditionally been a task for natural language processing and based explicitly on text data, such as online blog posts (Feng et al., 2011). Beyond the scope of text-based sentiment analysis, Chen et al. (1998) provides us with an early work on audio-visual emotion recognition and showed that bimodal classifiers can perform better than unimodal ones alone.

Even though there is a significant amount of research done on audio-visual emotion recognition, only a few previous efforts have systematically explored trimodal fusion by combining text data with audio and visual modalities. Morency et al. (2011) was one of the first to investigate sentiment analysis on video movie reviews. They analyzed a collection of 47 videos depicting monologues in addition to the corresponding text that they manually transcribed from each 30-seconds excerpt. They evaluated sentiment for each review as a 3-way classification problem: positive, negative or neutral and achieved an F1 measure of 55.3%, which is much better than chance.

Furthermore, Wöllmer et al. (2013) attempted the same type of multimodal sentiment task for movie reviews using a linear Support Vector Machines (SVM) for the linguistic features and a Bidirectional Long Short-Term Memory (BLSTM) for the audiovisual ones. Our work continues this direction of combining data from different modalities and we also used video movie reviews. However, these related studies used very small collections of videos, whereas our work uses more than 2,000 videos.

Poria et al. (2015) provided a novel use of deep Convolutional Neural Networks (CNNs). They extracted features from the text modality and then adopted multiple kernel learning (MKL) for classifying the multimodal fused feature vectors. Most previous work has verified that multimodal classifiers perform better than unimodal ones.

More recently, Poria et al. (2018) presented three fusion techniques for multimodal sentiment analysis which achieved high accuracy: concatenation-based fusion, context-aware fusion and context-aware fusion with attention. One major issue of early fusion is that input-level feature concatenation will increase the feature space, which can be potentially problematic for very large datasets. To account for this, we experimented with principle components analysis (PCA) as a dimensionality reduction technique.

Existing top-performing systems on the CMU-MultimodalDataSDK MOSI (Zadeh et al., 2018) dataset are listed in Table 1, measured by classification accuracy. The state-of-the-art is Zadeh et al. (2017) which used tenor-based multimodal fusion. The C-MKL system of Poria et al. (2015), as discussed earlier, used a novel approach with CNNs. We also include a non-DNN system from Zadeh et al. (2016b) because it used input-level feature fusion, similar to one of our approaches in this work. Note that each of these systems has used slightly different feature selection techniques, which have introduced some inconsistencies between systems making a direct comparison difficult. Thus, we cannot make a direct system-to-system comparison between our methods and previous work.Additional work has been carried out on unimodal and multimodal sentiment analysis, using datasets different from CMU MOSI (Poria et al., 2016; Ma et al., 2018).

System	Authors	Acc
TFN	Zadeh et al. (2017)	77.1%
GME-LSTM(A)	Chen et al. (2017)	76.5%
C-MKL	Poria et al. (2015)	73.1%
SVM-MD	Zadeh et al. (2016b)	71.6%

Table 1: Accuracy reported in previous work on trimodal fusion for binary sentiment classification using MOSI dataset. Note that these systems differ slightly in terms of data pre-processing.

3 Methodology

Here we provide the technical specifications of the DNN architectures and parameters that we used in this work, followed by details about our three fusion techniques. We then discuss PCA dimensionality reduction, which we used in our experiments

as a form of feature selection.

3.1 Data and Task Description

We conducted our experiments on the Multimodal Opinion level Sentiment Intensity (MOSI) dataset from CMU-MultimodalDataSDK (Zadeh et al., 2018).[1] The MOSI dataset is a collection of 2199 opinion video clips, each annotated with sentiment scores in the range [-3,3]: strongly positive (+3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3). The multimodal observations consist of transcribed speech and features extracted from the visual and audio data. This benchmark dataset provided pre-extracted features on three modalities, a speaker-independent data partition of train (1283 items), validation (229 items), and test (686 items) sets, and an alignment of text, acoustic and visual data.

A detailed description of the dataset features and the sentiment class labels can be found in Zadeh et al. (2018). We aligned the features to the text embeddings as a reference and we max-normalized the feature values on a per-modality basis, as this allows for a meaningful comparison across systems. Due to the different number of timesteps in each utterance, we were required to restrict each sentence to a fixed size length by padding or cropping the sentences, using a maximum length. We treated this maximum length as a hyper-parameter and is described in more detail.

Primarily, our prediction task is binary classification for sentiment: positive versus negative. An exemplar with score $s > 0$ belongs to the positive class, while scores of $s < 0$ belong to the negative class. We transformed all scores to True/False values, where True corresponds to the positive class. For performance metrics, we used overall accuracy on the held-out test set.

After we identified the best-performing overall trimodal fusion system, we conducted additional experiments to report 5-class accuracy with F1 measure, as well as regression where we report mean-absolute error (MAE). These additional metrics allow further comparison of our best system to existing systems for this dataset.

3.2 Unimodal classifiers

We describe three types of DNNs that we used in our experiments for sentiment prediction and some of the reasoning behind these selections.

Convolutional Neural Networks (CNNs) have been applied to various text-based sentiment and emotion detection tasks in natural language processing (Kim, 2014). Moreover, CNNs were used in OpenFace (Baltrušaitis et al., 2016), an open-source face recognition tool which was employed by MOSI. While there are limited studies that involve using CNNs to predict sentiment directly from speech, we note that others have successfully tested its efficacy by working directly on the speech spectrogram (Niu et al., 2017).

Long Short-Term Memory (LSTMs) are popular with sequence prediction tasks, because they can capture context from previous steps. LSTMs also achieved moderate success for video emotion detection Chen et al. (2017). We expect LSTMs to be useful in our sentiment prediction task due to the sequential nature of the video data.

Bidirectional LSTMs (BLSTMs) increase the amount of available contextual information by including both a forward pass and a backward pass through a sequence. There is growing interest in applying BLSTMs for emotion detection from visual and audio features (Ullah et al., 2018).

3.3 Training Hyper-parameters

The activation function we used across all of our experiments was ReLu (Nair and Hinton, 2010). The learning rule was Adam (Kingma and Ba, 2014) with standard parameters. For 1D convolution layers, the kernel size was 3. For max pooling layers, the window size was 2. We varied the number of convolutional layers in $[1, 2, 3]$. For LSTMs and Bi-directional LSTMs, we set the number of units to $[64]$ and the number of layers in $[1, 2, 3]$. For fully connected layers, we set the number of units to 100 and explored the number of layers in $[1, 2, 3]$. We added dropout (Srivastava et al., 2014) between fully connected layers with dropout rate in $[0.1, 0.2]$. In all experiments, we used early stopping with the stopping criteria set to identify maximum validation accuracy and patience was set to 10. We varied the maximum length setting for the video segments in our dataset, known as *maxlen*, in $[15, 20, 25, 30]$. The experiments employed batch normalization with batch size set to $b = 64$ (Ioffe and Szegedy, 2015). Since it is a binary classification task, we use a single output unit with sigmoid activation. The loss function we use is binary cross-entropy. We

[1] https://github.com/A2Zadeh/CMU-MultimodalDataSDK

present test set results measuring overall accuracy.

3.4 Input-Level Feature Fusion

Input-level feature fusion (early fusion) refers to simply concatenating features from all the modalities, after they have been aligned and transformed to fixed size length. The concatenation is performed on the time step dimension. After input concatenation, the process follows a standard deep learning pipeline and we can apply different deep learning structures on top of the concatenated features. In this work, we tested CNNs, LSTMs and BLSTMs. We explored using one fully connected hidden layer and one output layer for the final prediction. In each case, we optimize the hyper-parameters of the DNN as described earlier.

We experimented with dimensionality reduction on a per-modality basis, prior to feature concatenation. This is motivated by our observation that many of the visual and audio features were zero valued. Thus, we attempt to identify the most important features using PCA. Our system architecture for input-level fusion with and without dimensionality reduction is displayed in Figure 1.

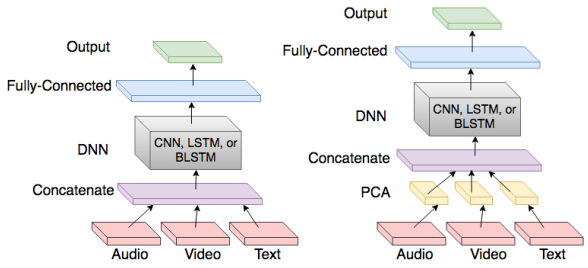

Figure 1: Input-level feature fusion architecture with and without PCA.

3.5 Intermediate-Level Feature Fusion

In intermediate-level feature fusion, data from each modality is first input to the best performing unimodal networks (for video and audio we use CNN, for text we use BLSTMs) which learn intermediate features. The intermediate weights from these unimodal networks are then concatenated and we then add fully connected layers to continue training the concatenated features. The goal is to capture interactions between modalities. We experimented with and without PCA on the input-level features. We show the architecture of the intermediate-level fusion system in Figure 2.

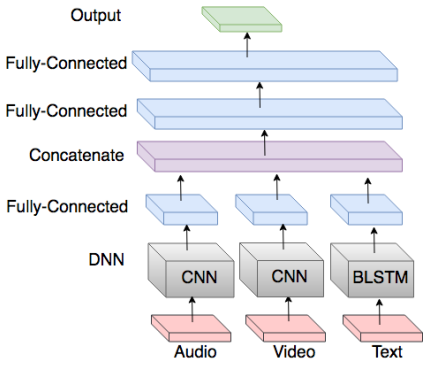

Figure 2: Intermediate-level feature fusion architecture. PCA for dimensionality reduction is not shown in this diagram.

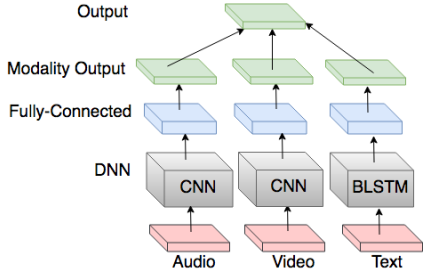

Figure 3: Late decision-level fusion architecture. PCA for dimensionality reduction is not shown in this diagram.

3.6 Decision-Level Feature Fusion

Decision-level feature fusion (late fusion) applies a separate classifier to weight the decisions of unimodal DNNs. The idea is that combining the unimodal results may improve model robustness. The most straightforward way of doing decision-level fusion is to train separate classifiers and weight their outputs with a tuple: $w = (\lambda_1, \lambda_2, \lambda_3)$. These weights can either be learned by another classifier, or set experimentally. No concatenation is performed in decision-level fusion. Compared to intermediate level fusion, which used sub-networks to extract intermediate features, here we output the decision of each modality.

Commonly, an SVM or another classifier is used on top of the decisions of each unimodal classifier. Our approach is different from existing literature in that we train 3 separate unimodal sub-networks such that our final system contains 3 component networks. For an illustration, refer to Figure 3. The top layer of this network is simply an output layer that receives the output of each modality sub-network (so the input is a one dimensional vector of size 3) and assigns a weight for

each. This architecture acts as an ensemble of the 3 separate modality classifiers. Although it is not the case for our experiments, it would be possible to pre-train each modality on a different dataset, if more data is available (Wu et al., 1999).

3.7 Principal Component Analysis (PCA)

We applied PCA as a way to select the most valuable features, and reduce the dimensionality of the feature space, and ultimately to increase the unimodal performance. Our goal in using PCA was to find the most effective and least redundant components to the unimodal representation of the data since features are semantically different after they are max-normalized (Zadeh et al., 2017).

PCA is an important linear transformation technique for dimensionality reduction. PCA yields the ordered feature vectors, commonly referred to as *principal components*, which maximize the variance of the data by removing redundant features (Abdi and Williams, 2010). As a data reduction technique, PCA is commonly used for handling high-dimensional visual information in various research areas, such as medical images (Bhat et al., 2017), and has been proved to be an effective method for feature selection and extraction.

We used the Python Sklearn PCA decomposition function (Pedregosa et al., 2011) on our training set. We computed the proportion of variance explained by the number of principal components utilized using a scree plot.[2] We then inferred a range of k-components that might be responsible for a high enough cumulative variance and swept this range of k-component values (shown in Table 2). We applied the PCA fit that we learned from training data and used it as the PCA transform on our validation and test data. We continued with the unimodal classifier training according to the fusion architectures and hyper-parameters described earlier. We then examined binary test accuracy on each DNN architecture to determine the best value for k in PCA. The top-performing system is highlighted in bold.

4 Experiment Results

In this section we provide the experiments on the 3 fusion techniques with and without PCA, for predicting the positive/negative sentiment of the videos. We report accuracy for the binary sentiment classification problem. After experimenting with the fusion techniques, we identify the best overall performing systems and further report the 5-class accuracy, F1, and regression MAE and correlation.

[2]commonly employed when there is a need to assess which components explain the most variability in the data, plots available upon request

DNN	Mode	Test Acc(%)		k, Var
		-PCA	+PCA	
LSTM	A	54.0	55.2	10, 0.61
BLSTM	A	53.0	55.1	10, 0.61
CNN	**A**	**55.2**	**57.2**	**20, 0.82**
LSTM	V	54.2	56.7	25, 0.94
BLSTM	V	55.8	56.5	20, 0.90
CNN	**V**	**57.8**	**57.1**	**25, 0.94**
LSTM	**T**	**70.1**	**71.7**	**110, 0.98**
BLSTM	T	69.7	70.8	110, 0.98
CNN	T	67.7	68.5	130, 0.99

Table 2: Unimodal binary accuracy, exploring k number of PCA components with corresponding variance threshold (A=audio, V=video, T=text).

4.1 Input-Level Feature Fusion

We explored input feature fusion with and without PCA. When we ran early fusion with PCA, we used the k-PCA components value described in Table 2. Our experiment results for early fusion are displayed in Table 3. The top-performing systems for each modality combination are highlighted in bold.

DNN	Mode	Test Acc(%)		Best
		-PCA	+PCA	Parameters
LSTM	A,V,T	70.5	70.1	1, 0.2, 25
BLSTM	**A,V,T**	**71.4**	**71.8**	**3, 0.2, 25**
CNN	A,V,T	69.2	68.5	1, 0.2, 20
LSTM	A,T	69.2	70.8	2, 0.2, 30
BLSTM	**A,T**	**71.2**	**71.2**	**1, 0.2, 25**
CNN	A,T	68.3	68.3	1, 0.1, 30
LSTM	V,T	72.3	69.5	2, 0.2, 30
BLSTM	**V,T**	**72.4**	**69.3**	**2, 0.2, 30**
CNN	V,T	69.3	68.8	3, 0.2, 30
LSTM	A,V	55.1	55.8	3, 0.1, 20
BLSTM	A,V	55.1	56.7	3, 0.1, 30
CNN	**A,V**	**55.6**	**57.4**	**2, 0.1, 30**

Table 3: Bimodal/trimodal binary accuracy for early fusion. Parameters refer to DNN layers, dropout rate, segment length.

The gains from PCA for input-level fusion are particularly small, which is counter-intuitive considering that early fusion concatenation increases the dimensionality of the data. The best-performing overall system was a BLSTM using bimodal text and video data at 72.4% binary accuracy without PCA. The CNN tends to perform less well across all bimodal/trimodal combinations, and this suggests that emotion prediction has a sequential aspect. That sequential aspect is picked up by the other DNNs that we tested.

4.2 Intermediate-Level Feature Fusion

The intermediate features fusion model we proposed adds dense layers on top of the intermediate weights extracted from each modality. There are other possible configurations to be explored, but we experimented with the simplest one. Compared to early fusion, the features for each modality are first fed to a different network. We have chosen the best performing network for each single modality as described in Table 2 (CNN for audio and video, and BLSTM for text) for the pre-fusion stages.

Mode	Test Acc(%)		Best
	-PCA	+PCA	Params
A,V,T	73.3	73.0	1, 0.1, 30
A,V	60.0	59.0	3, 0.1, 30
A,T	70.5	70.8	2, 0.2, 25
V,T	**74.0**	**74.0**	**3, 0.2, 30**

Table 4: Bimodal/trimodal binary accuracy for intermediate feature fusion. Parameters refer to DNN layers, dropout rate, segment length.

When we applied PCA for intermediate-level fusion, we applied it either to all modalities or none. This configuration makes it possible to make a direct comparison with our other approaches. Results are summarized in Table 4. We achieve our highest performance so far which was the bimodal fusion of video and text with binary accuracy of 74.0%. We note that this accuracy was achieved with and without PCA, suggesting either that our proposed system is robust to noise or that video and text data was not particularly noisy.

4.3 Decision-Level Fusion

For our decision-level fusion (late fusion) experiments, we kept the pre-fusion network consistent with intermediate fusion (CNN for audio and video, BLSTM for text). Experiment results are in Table 5. Our best result is for the trimodal inputs. We find that the results are not much different from a carefully trained text only predictor. Since the video and audio classifiers are much worse predictors than text. This indicates that a decision level classifier is not the best approach for the MOSI dataset. We noticed that the top-performing decision-level systems used less segment length context than our previous experiments, even though the performance is comparable. This could be due to the fact that the combination of modalities creates a form of information enhancement, so that less context is needed to make a prediction.

Mode	Test Acc(%)		Best
	-PCA	+PCA	Params
A,V,T	70.6	70.8	2, 0.1, 25
A,V	56.8	58.1	1, 0.2, 25
A,T	71.7	71.7	3, 0.1, 15
V,T	**72.5**	**72.0**	**1, 0.1, 30**

Table 5: Bimodal/trimodal binary accuracy for decision-level fusion experiments.

4.4 Detailed Top Performing Systems

To make a comparison to performance reported in previous work, we provide more specific performance metrics in Table 6, based on the top-performing systems from each of the 3 fusion methods that we have discussed. For each top system, we report the binary accuracy and $F1$ score, the 5-class accuracy, and the regression MAE and Pearson r correlation (values closer to $r = 1$ indicate positive correlation, while values closer to $r = -1$ indicate negative correlation).

All of our best-performing systems used bimodal (text+video) feature fusion instead of trimodal. Across all systems, we can generalize that leaving out the audio modality improved performance. Our top input-level fusion system ($Early$) was bimodal BLSTM without PCA. Our top intermediate-level fusion system ($Inter.$), was bimodal fusion regardless of PCA. Finally, our best decision-level system ($Late$) was bimodal without PCA.

5 Discussion and Analysis

Our unimodal experiments showed that applying PCA always yields improved performance for bi-

Top	Binary		5-class	Regress.	
Method	Acc	F1	Acc	MAE	r
Early	72.4	66.7	33.3	1.08	0.55
Inter.	**74.0**	**65.6**	**35.2**	**1.10**	**0.56**
Late	72.5	66.3	31.4	1.05	0.56

Table 6: Top fusion system performance on binary classification, 5-class classification and regression.

nary sentiment prediction on this dataset. Further, we were able to identify text as the single best-performing and audio as the worst-performing modality predictor. Although PCA improved unimodal performance, it did not have an effect on the intermediate and decision fusions. This could be due to inherent noise in the audio data from the CMU-MultimodalDataSDK, which our feature selection procedure did not remedy.

We present example negative and positive sentences in Table 7 and the scores given by our best performing classifier. A score above 0.5 classifies the sentences as positive. This outlines the difficulty of the task and shows that some sentences are difficult to label even for humans.

Sentence text	Truth	Score
The voice acting was phenomenal	+	0.94
It was like this like pouty like grumpy look	-	0.31
Now the real Steven Russel has like an IQ like 163 which is like wow genius	**+**	**0.49**
If you know they're in there this is a cheesy um movie	**-**	**0.80**

Table 7: Example sentences and their true labels. Incorrect classification is distinguished in bold/red.

6 Conclusions and Future Work

Despite our efforts to reduce feature redundancy during early fusion, we found an apparent ceiling in terms of the best binary accuracy, as it never reached above 74.0%. Our experiments showed that PCA improves test accuracy in the case of unimodal models and sometimes the early fusion model. Interestingly, in our bimodal and trimodal experiments, we found that leaving out audio and focusing on video+text features, always yields a slight improvement. This is consistent with the state of the art on the MOSI dataset (Zadeh et al., 2017) which found that audio is the weakest of all three modalities for this dataset. It would be interesting to disentangle whether or not this constitutes bias in the data or bias in human communication or perception of emotions.

As the goal of our study was to explore multimodal fusion techniques, we explored 3 different fusion architectures that all yield better results that unimodal classifiers. This indicates that there are interactions to be learned during the fusion process. We showed that both late decision-level fusion and early fusion can achieve comparable results. As a task for future work, we encourage exploring the best intervention point for intermediate-level fusion. For example, to vary the number of fully-connected layers on individual DNNs before concatenation. Similarly, it should be investigated how to weigh the DNNs before concatenation as we know that text is often the best unimodal predictor of sentiment.

In terms of combining the CNN architecture with PCA, CNNs will basically learn common structural components across the input features, which can be viewed as a redundancy that is removed by PCA. Therefore this combination would only useful to the extent that it helps with removing actual noise from the data. Similarly, this combination of CNN+PCA on audio-only data, which consists primarily of MFCC's, also creates a type of redundancy. Given that there could be better models than PCA, we encourage future work to systematically explore and compare techniques for both feature selection and noise reduction on the CMU MOSI dataset.

In the future, we plan to examine which of the low-level acoustic descriptors, facial features, and words are the most effective for sentiment analysis. This would help future studies to learn better feature representations for sentiment analysis. Further, we selected our top-performing models based on binary classification accuracy, without a category for "neutral". It could be the case that some of our data exemplars were a better fit for this third category, or that audio features are predictive of a neutral category, something that should be investigated in future work.

The MOSI dataset breaks down each movie review into sentences to be classified individually,

losing context that might be gained by looking at the other neighboring sentences. Motivated by Poria et al. (2017) who suggested contextual sentiment analysis, we plan on including additional contextual information when predicting the sentiment of a sentence. Instead of considering each utterance as a separate entity, we will add contextual information from neighboring sentences belonging to the same monologue and study the gain.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. The authors would like to thank Steve Renals at University of Edinburgh Centre for Speech Technology Research (CSTR) and the anonymous reviewers for their valuable comments.

References

Hervé Abdi and Lynne J Williams. 2010. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An Open Source Facial Behavior Analysis Toolkit. In *IEEE Winter Conference on Applications of Computer Vision*.

Mahima Bhat, Maya V Karki, et al. 2017. Feature Selection Based on PCA and PSO for Multimodal Medical Image Fusion Using DTCWT. *arXiv preprint arXiv:1701.08918*.

Lawrence S. Chen, Huang Thomas S., Tsutomu Miyasato, and Ryohei Nakatsu. 1998. Multimodal Human emotion/expression recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 366–371.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal Sentiment Analysis With Word-Level Fusion and Reinforcement Learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171. ACM.

Shi Feng, Daling Wang, Ge Yu, Wei Gao, and Kam-Fai Wong. 2011. Extracting Common Emotions From Blogs Based on Fine-Grained Sentiment Clustering. *Knowledge and Information Systems*, 27(2):281–302.

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, pages 448–456.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *CoRR*, abs/1408.5882.

Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Yukun Ma, Haiyun Peng, Tahir Khan, Erik Cambria, and Amir Hussain. 2018. Sentic LSTM: A Hybrid Network for Targeted Aspect-Based Sentiment Analysis. *Cognitive Computation*, pages 1–12.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 169–176, New York, NY, USA. ACM.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA.

Yafeng Niu, Dongsheng Zou, Yadong Niu, Zhongshi He, and Hua Tan. 2017. A Breakthrough in Speech Emotion Recognition Using Deep Retinal Convolution Neural Networks. *arXiv preprint arXiv:1707.09917*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-Level Multimodal Sentiment Analysis. In *Proceedings Empirical Methods in Natural Language Processing (EMNLP)*, pages 2539–2544.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 439–448. IEEE.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Amir Hussain, and Alexander Gelbukh. 2018. Multimodal Sentiment Analysis: Addressing Key Issues and Setting up Baselines. *ArXiv e-prints*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. 2018. Action Recognition in Video Sequences Using Deep Bi-Directional LSTM With CNN Features. *IEEE Access*, 6:1155–1166.

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems*, 28(3):46–53.

Lizhong Wu, S. L. Oviatt, and P. R. Cohen. 1999. Multimodal Integration - A Statistical View. *IEEE Transactions on Multimedia*, 1(4):334–341.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. *CoRR*, abs/1707.07250.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-Attention Recurrent Network for Human Communication Comprehension. In *arXiv preprint arXiv:1802.00923*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016a. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *CoRR*, abs/1606.06259.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016b. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6):82–88.

Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.